POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

CLARIN-PL
Common Language Resources and Technology Infrastructure

Speech Analysis in the Clarin-PL Project

Krzysztof Marasek, Danijel Koržinek, Mariusz Kleć, Łukasz Brocki, Krzysztof Wołk

Polish-Japanese Academy of Information Technology

Introduction

The contribution of PJAIT within CLARIN-PL is the development of tools that help Humanities and Social Sciences (HSS) researchers in deriving maximum information from their speech corpora, existing in the form of audio. Processing and analyzing this kind of data is conceptually difficult and can be expensive. Our aim is to connect technologies and modern algorithms with humanities data in the context of speech analysis. This connection requires user-friendly and user-oriented tools with easy access to the most relevant services. As a response to this need, currently we are focusing on creating a general purpose web service with user-oriented interface, which integrates the tools created thus far.

http://mowa.clarin-pl.eu : Polish speech tools
http://clarin-pl.eu : CLARIN-PL website

Polish Speech Tools and Services

The current **http://mowa.clarin-pl.eu** website provides the following speech services (some of them have not public user interface yet)

- **Grapheme-to-Phoneme** conversion of text from orthographic into phonetic form using SAMPA and IPA phonetic alphabet
- **Speech alignment** that accurately aligns provided transcriptions and audio recordings at word and phoneme level.
- **Speaker diarization** for unsupervised segmentation of audio into portions spoken by unspecified individuals
- **Keyword detection** providing the list of keywords the system generates a list of likely occurrences and their location.
- **Voice activity detection** for distinguishing parts of audio containing speech from those containing noise and other sources
- **Speech transcription** for converting speech recordings directly to text. It is based on the Kaldi toolkit for speech recognition and works best in specified domains.

Use Case: Polish Film Chronicles

The aim was to create a speech recognition system that is capable of recognizing narrator's speech in the corpus of Polish Film Chronicles (PKF) spanning the years 1945-1962. The dataset consisted of 4373 news segments amounting to 88 hours of audio. The advantages of this use case include fairly well defined domain and uniform quality. The disadvantages include low quality, noise, disruptions, high level of background music and lots of named entities. The corpus was semi-automatically corrected using the system which finally achieved WER of 10.97% (compared to 17.96% by Google).

Docker Integration

The pre-trained tools are available in the form of Docker images downloadable from:

```
https://hub.docker.com/u/danijel3/
```

More detailed information available on:

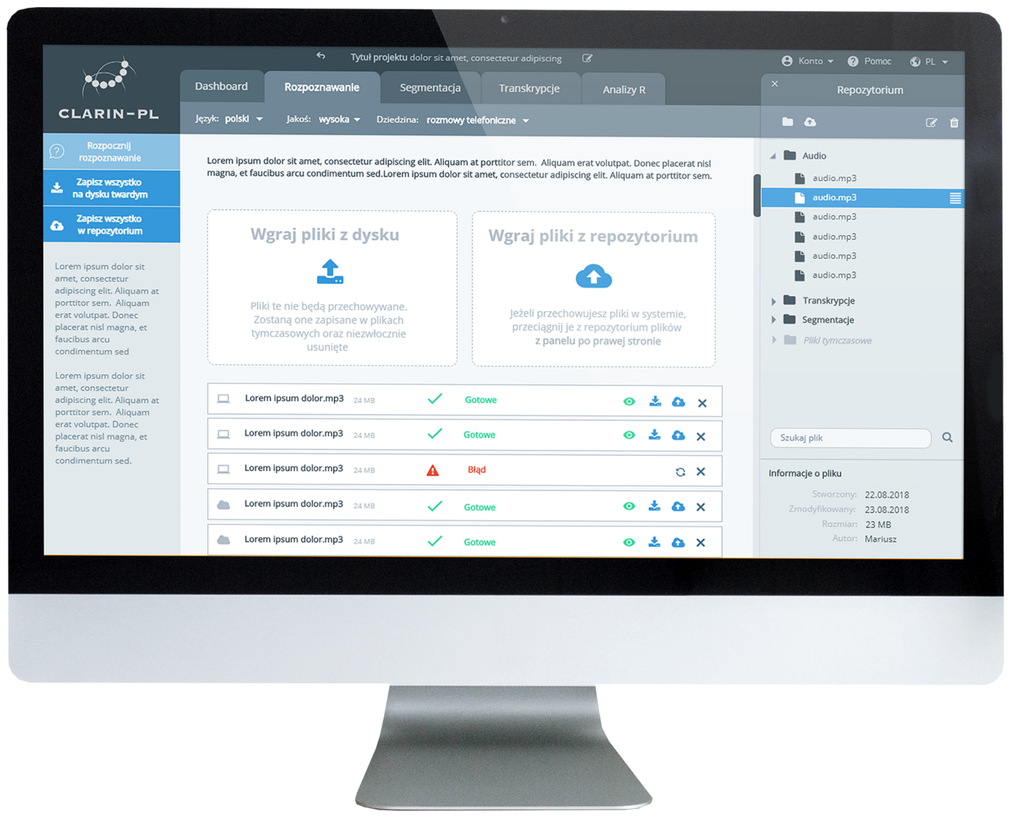
```
https://github.com/danijel3/SpeechToolsWorkers/tree/
master/speech_tools
```

Example of performing ASR on pkf.wav into trans.txt:

```
docker run --rm -v ./local:/data danijel3/clarin-pl-speechtools:pkf "/tools/Recognize/run.sh pkf.wav trans.txt"
```

Website Redesign

A redesign was initiated following the increased demand from the users. The prototype is the result of analysis the previous experiences, the flow of information and tasks in the existing system. It follows user-oriented practices in web design. The functional requirements documentation has been specified to support the future development of the website.



The key new features of the newly created prototype include:

- User friendly and easy to use interface based on drag and drop
- Central file repository with easy access and file management.
- Speech recognition of multiple audio files at the same time
- Speech alignment of multiple audio and and text files at the same time
- Strong help system including hints at every step of processing
- Ability to create and share multiple projects between users
- A tool for performing manual transcriptions
- Some necessary tasks work in background (e.g. normalization)
- Possibility to process the files without uploading them unto server

Future Plans

- Additional corpora will be annotated and delivered
- Integration with the EMU and Octra web applications
- Facilitation of cooperation with more partners in the HSS community
- Speech analysis directly in the web-browser

Contact Information

- **danijel@pjwstk.edu.pl - Danijel Korzinek**
- **mklec@pjwstk.edu.pl - Mariusz Kleć**
- kmarasek@pjwstk.edu.pl - Krzysztof Marasek
- lukas@pjwstk.edu.pl - Łukasz Brocki
- kwolk@pjwstk.edu.pl - Krzysztof Wołk